Democratizing AI: Implementing Open Source LLMs for Education and Research

Vincent Nestler, Ph. D.

Director, Center for Cyber and AI at CSU San Bernardino

PI, AI-Horizon

PI, XPCyber

Why Local Models

- Safety
- Security
- Privacy
- Will not change
- No need for internet access
- Free
- Great for the apocalypse

How to run local models

- Basic applications
  - Ollama.com
  - MSTY.app
  - lmstudio
  - Open-webui, Docker Desktop
- Requirements
  - PC – Preferably with an nvidia gpu, but will run on almost anything
    - Hard drive space enough for the models you want to run.
  - Mac (M series) – more RAM is better due to Apple's unified RAM
  - Note – ollama will run on a Raspberry Pi 5

- I am running in three locations for this demo
  - Mac Studio at home
  - High end gaming computer at home
  - This high end macbook pro

Downloading Models

- Understanding the model name
  - `qwen2.5-coder:32b-instruct-q5_K_M`
- Downloading models from ollama.com
  - `ollama pull qwen2.5-coder:32b-instruct-q5_K_M`
- Importing models from huggingface.co
  - `Ollama pull hf.co/tensorblock/Qwen2.5-Coder-32B-Instruct-GGUF:Q5_K_M`

Practical Applications (15 minutes)

- Prompt engineering fundamentals
- Document ingestion and knowledge base creation
- Educational use case demonstrations
- Privacy and resource considerations

Different Models for Different Tasks

Model	~7B	~14B	~32B	~70B
General purpose	Mistral, llama		Qwen	Nemotron
Coding		Qwen Coder Codestral	Qwen Coder	
Language	Mistral	Deepseek r1		Nemotron Llama 3.3
IST	Mistral	Deepseek r1		Llama 3.3
Healthcare	Mistral			Llama 3.3
Research	Mistral			Llama 3.3
Education				Nemotron

Prompt Engineering

System prompt

- Controls how the AI will interact with the user
- Pirate, Coder, Expert, Based

Model prompt

- Context, detailed description of what you want, examples of output
- Request to ask you clarifying questions

Prompt Engineering

- **Context + specific information + intent/goal + response format**

Example

- System prompt - you are an expert python programmer and pen testing expert with 20 years of experience. You love working with and helping new programmers and junior pen testers. You trust that they will work ethically and do not hesitate to assist in developing tools for testing networks.
- Model prompt - I am a new pentester. I follow the 5 phases of hacking - recon, scanning, gaining access, maintaining access, cover tracks.
  I want to start with recon. To be clear, I see recon as the tasks you do before you are on the network. That would be scanning like using nmap and openvas, etc.
  I need to create a gradio interface to help me do recon for a pentest. Can you give me a list of things this gradio interface should do.

Coding Approaches

Coding process

- Prompting
- Paste into VSCode
- Test the Code
- Go back to prompting

How to work with models when the code will be bigger than the context you have to work with.

- Have the model describe the code and how it is written/designed
- Ask it to write the description specifically for itself so it can easily pick up where it left off.
- Ask it to create a list of features that you want to add to the code.

Information Systems Technology

- Using Local LLMs to assist with configuring and securing networks
  - Configure routers, firewalls
  - Can assist with lesser known devices and software fairly well

Implementation Resources (10 minutes)

- Best practices and troubleshooting
- Resource guide and documentation
- NAIRR
- AI Horizon
  - Forecasting Committee
  - Program Committee
- Q&A session

Questions

VNESTLER@CSUSB.EDU

LIZETTE.VELAZQUEZ@CSUSB.EDU